

Image stitching network based on large-kernel feature perception and adaptive confidence gating

SONG ZHANG¹, ZHILIANG ZHAO^{2,*}, JIEHAO ZHAO³

¹ School of Electrical Engineering and Electronic Information, Xihua University, Chengdu 610039, China

² School of Aeronautics, Astronautics and Intelligent Equipment, Xihua University, Chengdu 610039, China

³ School of Electrical Engineering and Electronic Information, Xihua University, Chengdu 610039, China

*Corresponding author: zzl_caep@163.com

Image stitching is a key computational imaging technique for expanding the field of view of optical systems and generating high-resolution panoramas. However, eliminating ghosting artifacts and geometric distortions remains challenging under non-ideal imaging conditions involving large parallax or weak textures. To address these issues, this paper proposes a deep learning image stitching network based on large-kernel feature perception and adaptive confidence gating. First, ConvNeXt-Tiny is adopted as the feature extraction backbone to capture long-range geometric dependencies and global context. Second, an adaptive confidence gating mechanism dynamically suppresses erroneous mesh deformations in uncertain regions to ensure precise geometric alignment. Finally, an attention-guided fusion network optimizes feature alignment and suppresses artifacts. Experimental results on the UDIS-D dataset demonstrate that this method outperforms existing state-of-the-art approaches, achieving a 6.44% improvement in structural similarity index (SSIM) to 0.892 and a 0.13 increase in peak signal-to-noise ratio (PSNR) to 25.55. This approach effectively resolves parallax artifacts in image stitching and significantly enhances the structural consistency of panoramic images.

Keywords: image stitching, deep learning, computational imaging, geometric alignment.

1. Introduction

The trade-off between high resolution and a wide field of view (FOV) is a fundamental challenge in optical imaging system design [1]. While wide-angle and fisheye lenses can significantly expand the field of view, they often introduce severe edge distortion and reduced resolution [2, 3]. As an alternative, image stitching technology combines multiple images with overlapping fields of view into a seamless, high-resolution pano-

ramic image with a wide field of view [4]. This approach is promising for fields such as autonomous driving, virtual reality [2], robotic navigation [5], and optical remote sensing monitoring [6]. However, in practical applications, the optical centers of multi-camera systems or a moving camera typically do not coincide, resulting in complex geometric parallax between imaging planes. This poses significant challenges for achieving high-precision image registration.

Traditional image stitching methods [7, 8] typically assume that scenes lie on the same plane or that cameras undergo only rotational motion, using a single global homography for alignment. These approaches have achieved considerable success in both content alignment and shape preservation [8]. However, in practical applications, significant parallax often exists between input image pairs due to camera position shifts and variations in scene depth [9]. A single global transformation cannot accurately capture complex geometric distortions, leading to ghosting and misalignment in overlapping regions. To address this geometric optics challenge, subsequent research proposed mesh-based deformation methods, such as the traditional dual homography warping (DHW) model [10] and the as-projective-as-possible (APAP) warping model [11]. These approaches approximate non-planar transformations by partitioning images into multiple local regions. However, these methods rely heavily on feature point matching to determine transformation relationships and employ techniques like direct linear transformation (DLT) to compute spatial transformation matrices [11, 12], which often prove inadequate in low-texture, low-illumination scenes.

In recent years, deep learning-based methods have made significant progress in addressing complex scene stitching by extracting high-level semantic features from images [13]. Early work by DETONE *et al.* introduced the deep homography network (DHN) for homography estimation using supervised learning [14]. JAPKOWICZ *et al.* [15] improved upon this with a hierarchical stacked twin network. ZHANG *et al.* [16] employed residual networks with content masks to enhance homography estimation for dynamic objects. Since capturing perfectly aligned ground-truth in dynamic large-parallax scenes is physically impractical, unsupervised frameworks emerged. NIE *et al.* [13] framework advanced the field by modeling stitching as a dense mesh deformation problem. CHEN *et al.* [17] proposed an optimized seam-driven image stitching method incorporating scene depth, color, and texture information into a new energy function to reduce structural distortion near seams. ABHE [18] proposed leveraging attention mechanisms for precise homography estimation. DunHuangStitch [19] provided an unsupervised framework for stitching murals, while MGHE-Net [20] applied Transformers for multi-grid homography estimation. TANG *et al.* [21] explored robust stitching methods with straight-line preservation capabilities. Despite significant progress in these approaches, most existing networks suffer from limited receptive fields when handling extreme scenarios such as large parallax or weakly textured environments. This hinders the capture of long-range geometric dependencies, often leading to geometric distortions in high-parallax scenarios and limiting further improvements in image quality.

Based on the above analysis, this paper proposes an image stitching network that uses large-kernel feature extraction and multi-scale geometric alignment. First, to en-

hance the network's perceptual capabilities, we adopt ConvNeXt-Tiny as the encoder backbone. Through its 7×7 large convolution kernel, ConvNeXt-Tiny achieves a Transformer-like large receptive field while maintaining convolutional efficiency, enabling the extraction of more discriminative geometric features in baseline scenarios. Second, we propose an adaptive confidence-gated mechanism. This mechanism dynamically adjusts mesh deformation based on pixel-level matching reliability, adaptively suppressing invalid displacement estimates in high-uncertainty regions. This effectively prevents topological collapse of the mesh in non-significant areas.

2. Algorithm design

To address artifacts and geometric distortions in large-parallax image stitching, this paper proposes an image stitching network that uses large-kernel feature perception and adaptive confidence gating. As shown in Fig. 1, the proposed architecture integrates a collaborative perception registration network with an attention-guided fusion module. To achieve precise alignment, a weight-sharing ConvNeXt-Tiny backbone extracts multi-scale features, which are then processed through a coarse-to-fine strategy. Specifically, a global homography branch determines the initial global projective transformation, while a local mesh branch, enhanced by an adaptive confidence gating mechanism, predicts dense offset fields (parameterized by a 12×12 spatial grid) for warped image generation. To eliminate residual pixel-level misalignments at the boundaries, the second stage uses an attention-guided fusion network to produce high-quality masks, ultimately synthesizing a seamless panoramic image.

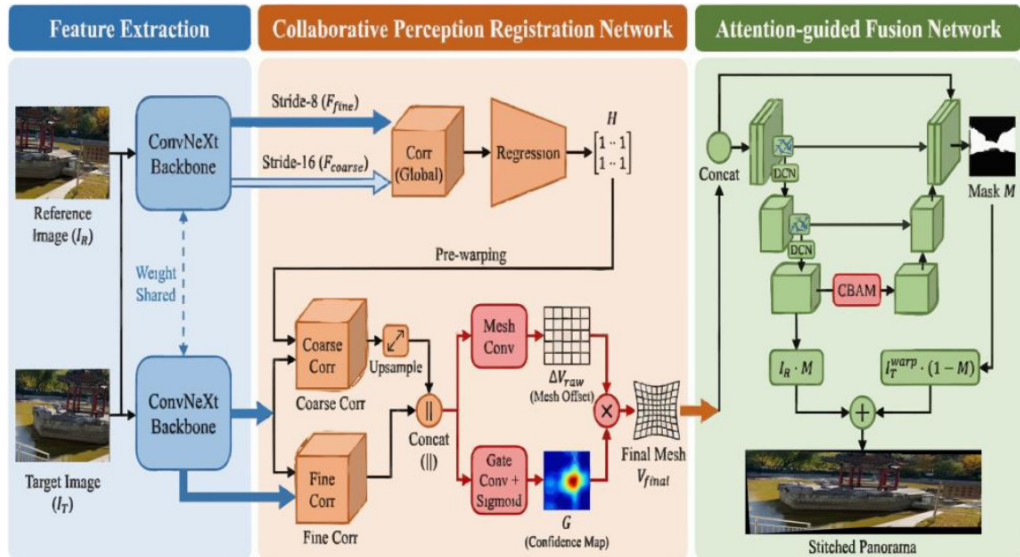


Fig. 1. The overall architecture of the proposed network. It consists of a collaborative perception registration module and an attention-guided deformable fusion module.

2.1. Global-to-local collaborative perception registration network

To overcome the limitations of restricted receptive fields and single-scale matching in existing methods, we construct a collaborative perception registration framework, synergizing global homography constraints with local mesh deformations, to achieve high-precision geometric alignment ranging from global structures to local details. Specifically, hierarchical features are simultaneously extracted from both the reference image I_R and the target image I_T using a weight-shared ConvNeXt-Tiny feature extractor. As shown in Fig. 2, features are extracted at two distinct scales to balance the receptive field and resolution.

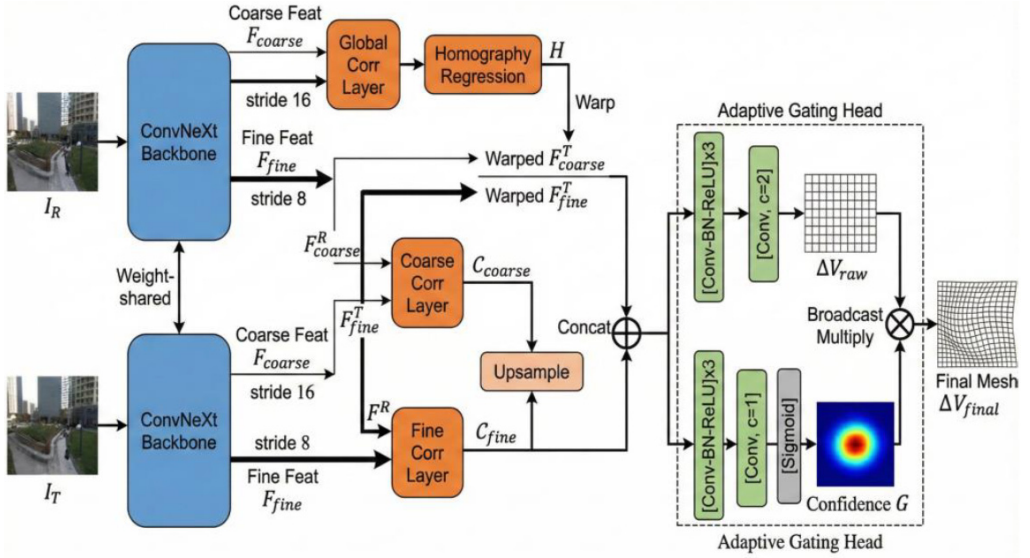


Fig. 2. Schematic illustration of the global-to-local collaborative perception registration network.

First, the global correlation layer and regression module estimate the global homography matrix H using coarse-grained features. Then, H is used to pre-deform the target features, roughly aligning them with the reference features. Subsequently, the content-aware correlation layer CCL [22] computes coarse-grained and fine-grained feature correlation volumes between the reference features and the pre-deformed target features:

$$C_{\text{coarse}} = \text{CCL}(F_{\text{coarse}}^R, w(F_{\text{coarse}}^T)) \quad (1)$$

$$C_{\text{fine}} = \text{CCL}(F_{\text{fine}}^R, w(F_{\text{fine}}^T)) \quad (2)$$

where w denotes a pre-deformation operation based on the global homography estimated in the previous stage, and F represents multi-scale features. To synthesize global guidance and local details, the coarse-grained correlation volume C_{coarse} is upsampled

and concatenated with the fine-grained correlation volume C_{fine} in the channel dimension. This fused representation is then fed into the regression head. In areas with weak or repetitive textures, forcing the network to regress dense mesh offsets often leads to severe geometric distortion. To mitigate this issue, a dual-branch adaptive gated head was designed, enabling the network to simultaneously predict both the original mesh offsets and a positional degree map G . Specifically, the concatenated features pass through convolutional layers and a Sigmoid activation to generate $G \in (0, 1)$, which dynamically evaluates pixel-level matching reliability. The final mesh deformation is then adjusted to:

$$\Delta V_{\text{final}} = \Delta V_{\text{raw}} \cdot \sigma(G) \quad (3)$$

where ΔV_{final} denotes the final modulated mesh offset, ΔV_{raw} denotes the initial predicted mesh offset, containing horizontal and vertical displacement vectors for each mesh vertex, operation denotes broadcast element-wise multiplication, enabling single-channel confidence to reweight spatial displacements in the X and Y directions. $\sigma(\cdot)$ is the Sigmoid activation function, which maps the confidence map G to the range $(0, 1)$.

2.2. Image fusion network

Although the registration network has corrected most geometric distortions, minute pixel-level misalignments may still persist in overlapping regions. To generate seamless and natural panoramas, we employ a lightweight fusion module as a post-processing step, as shown in Fig. 3. This network adopts the U-Net [23] encoder-decoder architecture. It takes the stitched reference image I_R and the deformed target image I_T^{warp} as inputs and predicts a pixel-level fusion mask M .

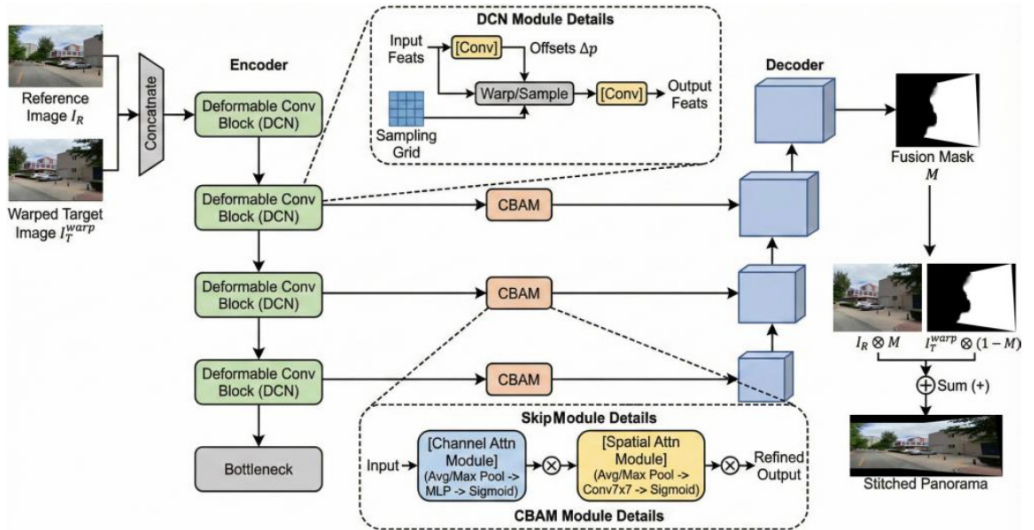


Fig. 3. Detailed structure of the attention-guided deformable fusion network.

In the standard U-Net, skip connections directly transmit features from the encoder to the decoder. However, these features may contain irrelevant background noise or misaligned artifacts. To suppress artifacts in non-overlapping regions and highlight salient features in overlapping regions, we integrate deformable convolutions (DCN) [24] with convolutional block attention modules (CBAM) [25] into the network. The final stitched panoramic image is synthesized via linear weighting according to the predicted mask:

$$S = I_R M + I_T^{\text{warp}}(1 - M) \quad (4)$$

where S represents the final stitched panoramic image, while $M \in [0, 1]$ denotes the learnable fusion mask predicted by the decoder.

Unlike traditional methods, this network learns to generate optimal weights. In non-overlapping regions, M tends toward binary values (0 or 1) to preserve original content. In overlapping regions, M generates smooth transitions to seamlessly blend textures, effectively concealing ghosting artifacts and exposure differences caused by misalignment.

2.3. Loss function design

Since obtaining the true deformation field for large-parallax scenes is impractical, we adopt a staged unsupervised training strategy. For the collaborative perception registration network, the objective function comprises pixel-level reconstruction loss and geometric smoothing loss. To drive the network to align the target image with the reference image, we impose photometric consistency constraints in overlapping regions. We adopt the L1 norm because it is more robust to intensity outliers and lighting variations compared to the L2 norm. The reconstruction loss L_{rec} is defined as

$$L_{\text{rec}} = \|I_R - I_T^{\text{warp}}\|_1 \quad (5)$$

Here, $\|\cdot\|_1$ denotes the L1 norm distance, while I_T^{warp} represents the deformed target image resulting from the prediction mesh deformation. This term encourages the deformed target to visually match the reference within shared regions.

To prevent mesh folding and preserve the topological regularity of the image, we introduce smoothing constraints on the predicted mesh vertices. We minimize the squared L2 differences between displacement vectors at adjacent mesh vertices to ensure the mesh does not undergo abrupt changes. The smoothing loss L_{smooth} is formulated as

$$L_{\text{smooth}} = \sum_{h, w} \left[\|u_{h+1, w} - u_{h, w}\|_2^2 + \|u_{h, w+1} - u_{h, w}\|_2^2 \right] \quad (6)$$

where $u_{h, w}$ denotes the predicted displacement vector of the mesh vertex at grid coordinate, while $u_{h+1, w}$ and $u_{h, w+1}$ represent the displacement vectors of its adjacent vertices in the vertical and horizontal directions, respectively.

The final total loss function L_{total} is a weighted sum of the reconstruction loss and the smoothing loss:

$$L_{\text{total}} = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{smooth}} L_{\text{smooth}} \quad (7)$$

where λ_{rec} and λ_{smooth} are the balancing weights. In the experiment, we set $\lambda_{\text{rec}} = 1$ and $\lambda_{\text{smooth}} = 8$ to prioritize the preservation of geometric structure while maintaining alignment accuracy.

3. Experiments and results analysis

3.1. Dataset description

The proposed algorithm was evaluated on the UDIS-D [13] dataset, a widely used benchmark for large-disparity scene image stitching in deep learning. The dataset comprises 10,440 training image pairs and 1,106 test pairs, spanning diverse, complex environments such as indoor rooms, urban streets, and natural outdoor landscapes. To adapt the network input and balance computational efficiency, we uniformly resized the input image pairs to 512×512 pixels during preprocessing.

3.2. Training strategies and details

This image stitching framework comprises two components: a collaborative perception registration network and an image fusion network. First, the registration network is trained independently to learn geometric deformations. The objective is to minimize alignment errors between the reference image and the deformed target image. Once the registration network converges, its parameters are frozen, and the fusion network is trained. This phase focuses on learning optimal synthetic masks to optimize seam lines and eliminate artifacts.

The entire experiment was implemented using the PyTorch [26] deep learning framework on a workstation equipped with an NVIDIA RTX 3090 GPU. During the image registration phase, we trained the model end-to-end in an unsupervised manner for 100 epochs. Parameter updates were performed using the AdamW [27] optimizer with an initial learning rate of $1e-4$ and weight decay set to $1e-4$. The learning rate was adjusted using the cosine annealing [28] algorithm. The batch size was set to 8. During the fusion stage, the fusion network was trained independently for 50 epochs with a batch size of 1. This stage used the Adam [29] optimizer with an initial learning rate of $1e-4$ and an exponential decay strategy with a decay factor of 0.97.

3.3. Evaluation indicators

To rigorously evaluate the geometric alignment performance of the proposed method, quantitative assessments are conducted on the output of the registration network. Specifically, we compute metrics within the valid overlapping region, which is dynami-

cally defined by the predicted mesh deformation mask due to the lack of ground-truth. Peak signal-to-noise ratio (PSNR) and structural similarity [30] (SSIM) are adopted as quantitative evaluation metrics to assess geometric alignment accuracy. PSNR evaluates luminance consistency using mean squared error (MSE). SSIM assesses structural degradation and geometric fidelity of the registered images. PSNR is measured in dB, while SSIM is a dimensionless index ranging from 0 to 1.

3.4. Algorithm comparison and analysis

To validate the effectiveness of the proposed algorithm, we conducted a comprehensive quantitative and qualitative comparison of this method with representative image stitching approaches. The comparison benchmarks included traditional feature-based methods (such as SIFT [31] and APAP) as well as advanced deep learning methods. Here, $I3 \times 3$ refers to the identity matrix used as a reference for a non-registration transformation. Based on performance, the evaluation results were divided into three sections. As shown in Tables 1 and 2, our method achieves the best performance across

T a b l e 1. Comparison of PSNR (dB) values for overlapping regions in the matching results of different methods for testing images from the UDIS-D dataset. The best results are highlighted in bold.

Algorithms	PSNR(↑)			
	Easy	Moderate	Hard	Average
$I3 \times 3$	15.82	12.72	10.66	13.07
SIFT+RANSAC	28.71	23.77	16.59	23.02
APAP	27.96	24.38	20.21	23.77
DHN	15.42	12.35	10.65	12.71
UDHN	18.25	14.02	11.11	14.41
UDIS2	30.18	25.82	21.55	25.42
Ours	30.39	25.93	21.64	25.55

T a b l e 2. Comparison of SSIM values for overlapping regions in the matching results of different methods for testing images from the UDIS-D dataset. The best results are highlighted in bold.

Algorithms	SSIM(↑)			
	Easy	Moderate	Hard	Average
$I3 \times 3$	0.531	0.281	0.142	0.302
SIFT+RANSAC	0.915	0.833	0.632	0.779
APAP	0.902	0.831	0.684	0.792
DHN	0.421	0.157	0.072	0.192
UDHN	0.719	0.521	0.355	0.530
UDIS2	0.932	0.870	0.742	0.838
Ours	0.955	0.916	0.824	0.892



Fig. 4. Visual comparison of stitching results in challenging scenes with large parallax. From left to right: input images, APAP, DHN, UDHN, UDIS2, and our method. Our approach preserves straight-line structures and eliminates ghosting artifacts in overlapping regions.

all metrics. Compared to the baseline UDIS2 method, our model improves SSIM from 0.838 to 0.892 (a 6.44% increase) and PSNR from 25.42 to 25.55. This demonstrates that the proposed Co-sensing Registration Network effectively captures global geometric structures, thereby generating higher-quality panoramas.

Beyond quantitative metrics, visual perceived quality is also a crucial criterion for evaluating stitching performance. Figure 4 presents visual comparison results between our method and baseline approaches in challenging scenarios.

Traditional methods (*e.g.*, APAP) often fail in low-texture regions, resulting in noticeable ghosting and structural discontinuities. While the deep learning-based UDIS2 improves overall alignment, it still exhibits structural distortion or blurring when handling complex linear structures [32] or high-parallax foregrounds [33]. In contrast, our method not only preserves the integrity and smoothness of linear structures but also effectively eliminates seam artifacts in overlapping regions through a post-processing image synthesis module. This demonstrates our approach's capability to generate visually natural, texture-preserving high-quality panoramas while ensuring high-precision geometric alignment.

3.5. Ablation studies and analysis

To thoroughly investigate the effectiveness of each proposed component, we conducted comprehensive ablation studies. We designed four network variants, as shown in Table 3.

Baseline model A uses a standard ResNet-50 backbone with limited receptive field and alignment capabilities. Model B replaces this backbone with ConvNeXt-Tiny, boosting the initial SSIM from 0.838 to 0.878. This improvement confirms that ConvNeXt's large-kernel convolutions are crucial for capturing the long-range geometric dependencies required for large-disparity scenes. Furthermore, integrating an adaptive con-

T a b l e 3. Ablation study on the effectiveness of key components. Model A is the baseline (ResNet50); model B introduces the ConvNeXt-Tiny backbone; model C incorporates the adaptive gating mechanism. The proposed method (model D) achieves the best performance.

Model	ResNet50	ConvNeXt-Tiny	Adaptive gating	PSNR	SSIM
A (baseline)	+			25.42	0.838
B		+		25.49	0.878
C	+		+	25.44	0.843
D		+	+	25.55	0.892

fidence gating mechanism (model C) further improves SSIM from 0.838 to 0.843, validating that estimating matching uncertainty helps suppress erroneous deformations in ambiguous regions. Finally, the fully integrated collaborative perception alignment network (model D), which combines large-kernel feature perception with adaptive confidence gating, achieves the highest structural similarity index (SSIM) of 0.892 and a peak signal-to-noise ratio (PSNR) of 25.55 dB. These results demonstrate that integrating large-kernel feature perception with adaptive confidence gating enables high-quality image stitching in large-disparity environments.

3.6. Discussion

The superior performance of this method stems from its effective integration of large receptive field feature perception with an adaptive confidence gating mechanism. Traditional feature-based approaches (*e.g.*, SIFT [31]) struggle to extract reliable keypoints in environments with weak or repetitive textures, while early deep learning methods [13] are constrained by the small receptive fields of standard convolutions. We significantly expand the effective receptive field by using the ConvNeXt-Tiny backbone network. This enables the network to capture long-range geometric dependencies and global semantic context. Beyond enhanced global perception, the proposed adaptive confidence gating mechanism plays a crucial role in improving the structural consistency of stitched images. By learning matching uncertainties, it dynamically suppresses erroneous mesh deformation estimates from unreliable regions, preventing mesh topology collapse (*i.e.*, vertex overlapping or self-intersection [13]), a common issue in non-overlapping or texturally ambiguous areas. This approach ensures high-precision geometric correction while preserving the natural image structure.

4. Conclusion

This paper addresses the inherent challenges of large parallax and geometric distortion in wide-field optical imaging systems by proposing a robust image stitching network that uses large-kernel feature perception and adaptive confidence gating. By leveraging the ConvNeXt-Tiny backbone network, this method effectively establishes long-range geometric dependencies, overcoming the limited receptive field of traditional convo-

lutional networks. A core innovation of this framework is the integration of adaptive confidence gating. This module acts as a soft filter to suppress noise-induced deformations by explicitly evaluating ambiguity in feature matching within low-texture or repetitive-texture regions, thereby achieving precise geometric alignment. Extensive experiments on the UDIS-D dataset demonstrate the method's superiority, achieving a structural similarity index (SSIM) of 0.892 and a peak signal-to-noise ratio (PSNR) of 25.55 dB. These results validate our computational imaging approach's ability to overcome the physical trade-off between field of view and resolution in panoramic imaging. Future work will focus on optimizing the network architecture for real-time processing in embedded optical detection systems and extending this framework to multimodal applications (e.g., visible-infrared image stitching).

References

- [1] WANG C., MING G.Y., *Modeling and evaluation of performance of dual field-of-view common-aperture dual-band imaging system*, *Optica Applicata*, **52**(4), 2022: 535–550. <https://doi.org/10.37190/oa220405>
- [2] ANDERSON R., GALLUP D., BARRON J.T., KONTKANEN J., SNAVELY N., HERNANDEZ C., AGARWAL S., SEITZ S.M., *Jump: Virtual reality video*, *ACM Transactions on Graphics (TOG)*, **35**(6), 2016: 198. <https://doi.org/10.1145/2980179.2980257>
- [3] LI D., ZHANG B., ZHU J., WANG Q., ZHU Z., *A study of distortion correction algorithms based on aspheric fisheye lens design*, *Optica Applicata*, **50**(4), 2020: 579–592. <https://doi.org/10.37190/oa200406>
- [4] NIE L., LIN C., LIAO K., LIU M., ZHAO Y., *A view-free image stitching network based on global homography*, *Journal of Visual Communication and Image Representation*, **73**, 2020: 102950. <https://doi.org/10.1016/j.jvcir.2020.102950>
- [5] DEWANGAN A.K., RAJA R., SINGH R., *An implementation of multi sensor based mobile robot with image stitching application*, *International Journal of Computer Science and Mobile Computing*, **3**(6), 2014: 603–609.
- [6] MIHOUB Z., HASSINI A., *Monitoring and identification of marine oil spills using advanced synthetic aperture radar images*, *Optica Applicata*, **44**(3), 2014: 433–449. <https://doi.org/10.5277/oa140308>
- [7] BROWN M., LOWE D.G., *Automatic panoramic image stitching using invariant features*, *International Journal of Computer Vision*, **74**(1), 2007: 59–73. <https://doi.org/10.1007/s11263-006-0002-3>
- [8] CHEN Y.S., CHUANG Y.Y., *Natural image stitching with the global similarity prior*, [In] Leibe B., Matas J., Sebe N., Welling M. [Eds.], *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, Vol. 9909, Springer, Cham, 2016: 186–201. https://doi.org/10.1007/978-3-319-46454-1_12
- [9] LI L., YAO J., XIE R., XIA M., ZHANG W., *A unified framework for street-view panorama stitching*, *Sensors*, **17**(1), 2017: 1. <https://doi.org/10.3390/s17010001>
- [10] GAO J., KIM S.J., BROWN M.S., *Constructing image panoramas using dual-homography warping*, [In] *CVPR 2011*, Colorado Springs, CO, USA, 2011: 49–56. <https://doi.org/10.1109/CVPR.2011.5995433>
- [11] ZARAGOZA J., CHIN T.J., BROWN M.S., SUTER D., *As-projective-as-possible image stitching with moving DLT*, [In] *2013 IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, 2013: 2339–2346. <https://doi.org/10.1109/CVPR.2013.303>
- [12] HARTLEY R., ZISSERMAN A., *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge 2003.
- [13] NIE L., LIN C., LIAO K., LIU S., ZHAO Y., *Unsupervised deep image stitching: Reconstructing stitched features to images*, *IEEE Transactions on Image Processing*, **30**, 2021: 6184–6197. <https://doi.org/10.1109/TIP.2021.3092828>

- [14] DeTONE D., MALISIEWICZ T., RABINOVICH A., *Deep image homography estimation*, arXiv:1606.03798. <https://doi.org/10.48550/arXiv.1606.03798>
- [15] NOWRUZI F.E., LAGANIERE R., JAPKOWICZ N., *Homography estimation from image pairs with hierarchical convolutional networks*, [In] *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, Italy, 2017: 904–911. <https://doi.org/10.1109/ICCVW.2017.111>
- [16] ZHANG J., WANG C., LIU S., JIA L., YE N., WANG J., ZHOU J., SUN J., *Content-aware unsupervised deep homography estimation*, [In] Vedaldi A., Bischof H., Brox T., Frahm J.M. [Eds.] *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, Vol. 12346, Springer, Cham, 2020: 653–669. https://doi.org/10.1007/978-3-030-58452-8_38
- [17] CHEN X., YU M., SONG Y., *Optimized seam-driven image stitching method based on scene depth information*, *Electronics*, **11**(12), 2022: 1876. <https://doi.org/10.3390/electronics11121876>
- [18] Huo M., ZHANG Z., REN X., YANG X., YE C., *AbHE: All attention-based homography estimation*, *IEEE Transactions on Instrumentation and Measurement*, **73**, 2024: 5013911. <https://doi.org/10.1109/TIM.2024.3374320>
- [19] MEI Y., YANG L., WANG M., YU T., WU K., *DunHuangStitch: Unsupervised deep image stitching of Dunhuang murals*, *IEEE Transactions on Visualization and Computer Graphics*, **31**(8), 2025: 4226–4240. <https://doi.org/10.1109/TVCG.2024.3398289>
- [20] TANG Y., TIAN S., SHUAI P., DUAN Y., *MGHE-Net: A transformer-based multi-grid homography estimation network for image stitching*, *IEEE Access*, **12**, 2024: 49216–49227. <https://doi.org/10.1109/ACCESS.2024.3384598>
- [21] TANG Z., TANG Q., *Robust image stitching method with line preservation and seamless blending*, *Applied Optics*, **64**(29), 2025: 8682–8690. <https://doi.org/10.1364/AO.568424>
- [22] NIE L., LIN C., LIAO K., LIU S., ZHAO Y., *Depth-aware multi-grid deep homography estimation with contextual correlation*, *IEEE Transactions on Circuits and Systems for Video Technology*, **32**(7), 2022: 4460–4472. <https://doi.org/10.1109/TCSVT.2021.3125736>
- [23] RONNEBERGER O., FISCHER P., BROX T., *U-Net: Convolutional networks for biomedical image segmentation*, [In] Navab N., Hornegger J., Wells W., Frangi A. [Eds.], *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, Vol. 9351, Springer, Cham, 2015: 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
- [24] DAI J., QI H., XIONG Y., LI Y., ZHANG G., HU H., WEI Y., *Deformable convolutional networks*, [In] *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 764–773. <https://doi.org/10.1109/ICCV.2017.89>
- [25] WOO S., PARK J., LEE J.Y., KWEON I.S., *CBAM: Convolutional block attention module*, [In] Ferrari V., Hebert M., Sminchisescu C., Weiss Y. [Eds.], *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, Vol. 11211, Springer, Cham, 2018: 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
- [26] PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KOPF A., YANG E., DeVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J., CHINTALA S., *PyTorch: An imperative style, high-performance deep learning library*, [In] Wallach H., Larochelle H., Beygelzimer A., d’Alché-Buc F., Fox E., Garnett R. [Eds.], *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc., 2019.
- [27] LOSHCILOV I., HUTTER F., *Decoupled weight decay regularization*, [In] *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
- [28] LOSHCILOV I., HUTTER F., *SGDR: Stochastic gradient descent with warm restarts*, [In] *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.
- [29] KINGMA D.P., BA J., *Adam: A method for stochastic optimization*, [In] *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [30] WANG Z., BOVIK A.C., SHEIKH H.R., SIMONCELLI E.P., *Image quality assessment: From error visibility to structural similarity*, *IEEE Transactions on Image Processing*, **13**(4), 2004: 600–612. <https://doi.org/10.1109/TIP.2003.819861>

- [31] LOWE D.G., *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, **60**(2), 2004: 91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- [32] XUE W., XIE W., ZHANG Y., CHEN S., *Stable linear structures and seam measurements for parallax image stitching*, IEEE Transactions on Circuits and Systems for Video Technology, **32**(1), 2022: 253–261. <https://doi.org/10.1109/TCSVT.2021.3058655>
- [33] LIAO T., LI N., *Natural image stitching using depth maps*, Signal Processing: Image Communication, **141**, 2026: 117438. <https://doi.org/10.1016/j.image.2025.117438>

*Received February 2, 2026
in revised form March 4, 2026*